Weekly Report

Lu Junhua

2015 年 6 月 29 日

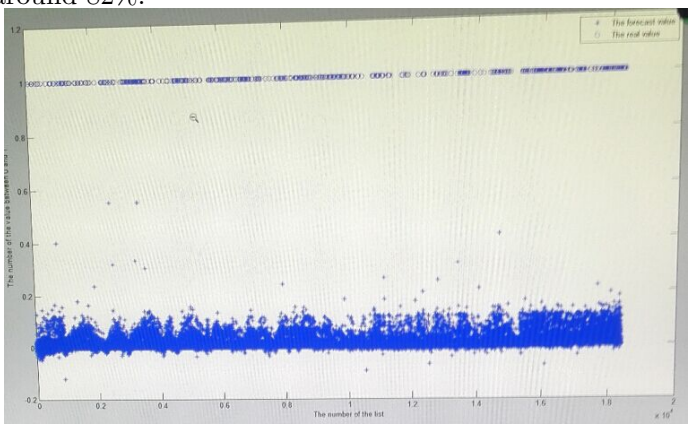
This week, I have extracted the data and made them into two samples.

- After consulting Gu tianyu, I re-encoded the occupation as three categories:00(Someone who has an occupation), 90(We know where he works, but are not clear about his job), 99(unemployed).
- As Prof. Gu said, I should compute the days of 开房per month/quarter. However, while I aggregate the data by month/quarter, I found that the records between June, July 2011 account for 99.7% of all the records. We have no choice but to aggregate all the time period. Here, I also notice that if someone changes his/her marital status, he/she will leave a record but without timestamp; if one temporary resident changes his residence will also leave a record. So there still many details in it, but I have to ignore it to proceed our study.
- We add attribute 省市区县 into the sample(consider it as one 6-digit number), and we do not do any regularization since Prof.Gu thought it is a GENERALIZED regression. All the attributes we used are: age (by month), 开房总时长(by day), marital status, 省市区县(身份证前六位), level of education(文化水平), gender, employment.
- Two samples, all the records are complete, all the records with one or two attributes missing are removed. permanent resident/temporary resident population is around 4:1.
- sample1:crime 524, innocent: 17872 sample2:c 430, i 18494.

	截距	年龄（月）	开房时长（天）	婚姻状况	户籍地	文化程度	性别	职业	备注
θ	-4.8269	1.0044*10^(-4)	0.0051	0.0049	2.8724*10^(-6)	0.0361	-1.3114	-0.0064	sample1
θ	-0.6944	-0.00082578	0.0033	0.0118	1.5783*10^(-7)	0.0345	-1.3233	-0.005	1572个数据点

图 1: logistic regression result

We use logistic regression and neural network in matlab. If we use the whole sample1 to regression, where most of the people did not commit a crime, it will return a result which is not satisfying. Input sample2 and we found almost no one commit crimes. So we randomly choose 1048 innocent in sample1, along with all the criminals in smaple1 to do regression and NN. Then we put sample2 dataset into the model, both model has a accuracy of around 82%.



So how can the accuracy be so high? Again, we may notice that most of the people are innocent. If we just use the criminals in sample2, the accuracy is 43.65%. And this result may reflect the true results of the regression.
We may do some more exploration next week.